Concepts for Hybrid Audio Coding Schemes Based on Parametric Techniques

Bernd Edler, Heiko Purnhagen

Laboratorium für Informationstechnologie
Uni Hannover, Schneiderberg 32, D-30167 Hannover, Germany
Tel.: +49 511 762 5035, Fax: +49 511 762 5052
E-Mail: `edler@tnt.uni-hannover.de`
URL: `http://www.tnt.uni-hannover.de/~edler/`

Abstract

Recently an audio coding technique based on a signal representation by parameters describing harmonic tones, individual sinusoidal components and noise components has been presented, which shows its highest efficiency at very low bit rates. In this paper concepts for combination with other techniques for extending the applicability at higher bit rates and for improving the quality in the presence of speech are presented.

1 Introduction

A recently presented object based analysis/synthesis audio coder for very low bit rates is based on a signal representation which describes different components of an audio signal by appropriate parameter sets [1]. Together with transform coding and speech coding techniques it has been integrated in the MPEG-4 Audio development [2]. For harmonic tones the most important parameters are the fundamental frequency and the amplitudes of the partials. Individual sinusoidal components are represented by their frequencies and amplitudes, while noise components are described by their spectral envelopes. Additional amplitude envelope parameters can be transmitted to improve the representation of transient signals. However, no phase information for the sinusoids is transmitted, since it is regarded not be relevant enough. While this technique shows high efficiency for coding of 4 kHz bandwidth music signals at bit rates between 4 and 8 kbit/s, the quality does not increase accordingly, if the bit rate is significantly increased. Furthermore at the lower end of this bit rate range and even below, specialized speech coders are available which provide higher efficiency in case of clean speech. Therefore two concepts for extensions of the parametric audio coder are presented. On one hand quality improvements at higher bit rates can be achieved by combination with transform coding techniques. On the other hand an improved efficiency for mixtures of speech and music signals can be obtained by combination with parametric speech coding techniques.

Section 2 gives a short introduction into the structure of the parametric audio coder and reviews the underlying models.

The first of the two concepts for extension mentioned above is to use coding tools which have been developed for combining a transform coder with a low rate speech or audio coder to form a bit rate scalable system [3]. These tools allow a selection in the frequency domain between encoding either the original signal or the residual, i.e. the difference of the original and decoded signal from the low rate coder. In order to combine the parametric audio coder with this technique, the generation of a decoder output signal minimizing the residual must be possible. Therefore Section 3 addresses the needs for transmission of additional information to extend the parameter set described above.

The second approach is to extend the parametric audio coder by combining it with specialized speech coding techniques. First proposals which had been presented for the combination of the parametric audio coder with a parametric speech coder operating at bit rates between 2 and 4 kbit/s [5] imposed the following restrictions onto the encoder [6]: either speech and music signals were assumed not to occur simultaneously or had to be available in separated signals. However there were no solutions for mixtures of speech and music signals. Section 4 presents an approach which is based on the extension of the object based analysis/synthesis technique as it is used in the existing parametric audio coder.

Conclusions and consideration for further development of hybrid coding schemes based on the object based analysis/synthesis audio coder are given in Section 5.

2 Overview on the Parametric Audio Coder

The general idea of the parametric audio coder - which is the basis of the investigations described in this paper - is to use an object-based source model for audio signals. In this approach it is assumed that the input signal is composed of different components which can be regarded as audio objects. Different types of source models with corresponding model parameters are available to describe these objects, since they may have different characteristics. In the following such a source model for an audio object is denoted as object model.

Currently the coder supports three different object models, i.e. individual sinusoids, harmonic tones and noise. For each frame of input samples the encoder separates the signal into objects corresponding to the three object models and estimates the parameters describing the specific characteristics of each object. The order of processing is such that first, if present, the parameters for a harmonic tone are extracted and a residual is calculated by using a signal synthesized from these parameters. Next the parameters for individual sinusoids are estimated and the residual is reduced iteratively by again subtracting synthesized components. The remaining signal is assumed to be colored noise and therefore parameters describing the spectral shape are estimated (see Figure 1).

After their extraction the parameters are quantized and coded under control of a psychoacoustic model specifically developed for parametric coding [7]. This control includes adaptive partitioning of the available data rate amongst the different objects, so that for example the number of individual sinusoids can be limited and only those components are transmitted which have the strongest influence on the subjective sound quality. In detail the parameters transmitted for the different objects are:

For a harmonic tone the fundamental frequency and the amplitudes of its partials are transmitted after quantization with logarithmic characteristic. If more than 10 partials have to be coded, starting from the 10th they are grouped with a group width increasing with frequency and only average amplitudes are transmitted.

For an individual sinusoid the parameters frequency and phase are transmitted after quantization with logarithmic characteristic. However, if a line is detected which continues from one frame to another, i.e. the frequency and amplitude changes are sufficiently small, the changes of these parameters are transmitted instead of the absolute values.

For both types of tonal signals no phase information is transmitted, although it is available in the encoder as a result of the parameter extraction.

For the noise object the shape of the spectral envelope is DCT coded. By selecting the number of transmitted DCT coefficients the accuracy of this representation can be selected.

The decoder (Figure 2) reconstructs the parameters and synthesizes the tonal signals by sine wave generators. For lines which continue from frame to frame phase continuity is provided. The noise component is generated by an inverse FFT on a spectrum with a magnitude corresponding to the spectral envelope and with random phases.

3 Combination with Transform Coding Techniques

Recently coding tools have been presented which allow the combination of a transform coder with another coder operating at a very low bit rate e.g. a CELP speech codec [3]. With this approach it becomes possible to combine the advantages of transform coding techniques like those of the MPEG-2 Advanced Audio Coding (AAC) with the benefits of other techniques optimized for another bit rate range or optimized for other types of input signals. Since these tools have been proven to be very efficient they have been accepted for scalable coding in the evolving MPEG-4 Audio standard.

In the encoder of such a scalable configuration first the encoder and decoder of a very low rate codec, generally denoted as core codec, has to be applied. This core codec even can operate at a lower sampling rate than the transform coder. In this case down-sampling is necessary prior to encoding. The output of the decoder is given to a transform which operates on the same sampling rate as the transform coder, which is applied to the original input signal after appropriate delay compensation. The transformed values are given to one of the scalability tools, i.e. the frequency selective switch which is inserted between the transform and the quantization units. For each of several predefined frequency ranges this switch can select either to pass the original spectral components of the transform coder or to calculate the differences. The criterion for this decision is to select the signal which can be coded more efficiently and the switching states need to be transmitted as side information (Figure 3).

In the scalable decoder the output of the core decoder is again given to the transform and to the frequency selective switch in the same way as it is done in the encoder. Controlled by the transmitted side information the switch can either pass the decoded spectral components from the transform coder or the superposition of both components to the inverse transform. If only the data from the core coder is available, it is still possible to just use the core decoder for obtaining an audio signal with reduced quality.

Using the parametric audio coder as a core in such a scalable configuration is only useful if the differences of the spectral components are likely to be smaller than the original spectral components. However this is usually not the case as long as there can be significant phase differences between the synthesized sinusoids and the corresponding components in the input signal. Due to the fact that there is no phase information transmitted, both even can be fully out of phase. Therefore so-called enhancement information needs to be transmitted for the parametric coder in addition to the normal parameters to enable a scalable system with a parametric core coder. Of course this information is not needed in cases where the parametric core decoder is used on its own. Therefore all the data only needed in a scalable decoder should be combined in a separate bit stream and transmitted with the data of the transform coder.

In the following the question is addressed what kind of information must be available to a parametric decoder core in order to significantly reduce the differences of spectral components. For the sake of simplicity first the parameter quantization for an individual sinusoid is considered.

The most obvious is the need to transmit phase parameters for the tonal components, i.e. the individual sinusoids and the partials of harmonic tones. Uniform quantization can be applied to these parameters which are available in the encoder from the parameter estimation. The number of bits needed for each phase value, i.e. for each sinusoidal per frame, can be derived from the maximum possible amplitude in the residual caused by a phase quantization error. For example by quantization with 5 bits the residual error for an individual sinusoid can be reduced to less than 10% of the original amplitude.

However also the frequency parameter quantization can have a significant influence on the maximum residual amplitude. For the normal operating mode of the parametric coder the required accuracy for frequency parameters only depends on the audibility of frequency deviations or frequency modulation like artifacts. Therefore a logarithmic quantizer characteristic is appropriate which results in a constant maximum relative deviation over the whole frequency range. Subjective evaluations resulted in quantizers with a maximum relative deviation of about 0.24%. However any frequency deviation leads to a phase deviation which increases with the time distance to the instant of correct phase. If sinusoids are synthesized with constant parameters for each frame, like it is done in the parametric audio coder, the maximum phase deviation can be calculated from the frequency deviation and the frame duration $\Delta\phi_{max} = \Delta f_{max} \cdot T_{frame} \cdot 180^\circ$ if the transmitted phase is valid for the frame center. Due to the relative maximum frequency deviation the maximum phase deviation is sufficiently low at low frequencies. For example at a frequency of 70 Hz and a frame duration of 32 ms the maximum phase deviation is about 1° leading to a maximum residual error amplitude of 1.6% of the original amplitude. If however the frequency of the sinusoid is 3000 Hz, the maximum phase deviation is 41° so that the residual amplitude is only reduced to 66% of the original amplitude.

These considerations show that for a scalable configuration uniform quantization of the frequency parameters would be advantageous, while for the stand-alone parametric coder the logarithmic characteristic is the optimum. This problem can be overcome by including frequency parameter refinement data to the enhancement information. The number of additional bits can be selected depending on the value of each frequency parameter, since the information of the stand-alone parametric core can be decoded first without the enhancement information. With the following assignment, the maximum phase deviation can be kept below 14°: 1 bit for frequencies between 1000 and 2000 Hz, 2 bits for frequencies between 2000 and 4000 Hz and so on.

A third problem arises for the partials of harmonic tones. On one hand their number can be very large, e.g. 100 for a fundamental frequency of 40 Hz and a bandwidth of 4 kHz, so that the enhancement information for all corresponding phase parameters would become too large. On the other hand only average amplitudes are transmitted for groups of partials above the 10th as described above. Therefore additional information would be required for the refinement of the individual amplitudes. A good compromise is to restrict the enhancement information to the phase parameters of the first partials with a maximum of 10 values. All partials for which no phase parameter is transmitted must not be included in the signal fed to the scalability tool.

Finally it should be noted that the synthesized noise component would cause an increase instead of a decrease of the residual signal. Therefore it also should not be added to the signal given to the scalability tool.

From the above considerations it can be seen that for the use of the parametric audio coder in a scalable configuration an encoder is needed which generates the additional enhancement information, i.e. phase parameters and frequency refinement parameters. Furthermore the decoder must support the following two modes: In the normal operating mode it is equivalent to the stand-alone decoder. In the mode which additionally processes the enhancement information, there are some modifications necessary. First it must be able to refine the frequency parameters and to synthesize sinusoids with defined phase parameters. Furthermore it must synthesize harmonics with a maximum of 10 partials and it must not synthesize a noise component.

In order demonstrate the influence of the parametric coder on the residual, Figure 4 shows the short time spectra of the original signal and of the residual for an excerpt of the test signal "eddie rabbit". It clearly shows the reduction of the first 10 partials of a harmonic tone with a fundamental frequency of about 244 Hz and of some additional individual sinusoids, especially of two strong components below 200 Hz.

One criterion for the evaluation of scalability tools is whether the obtained efficiency is higher than that in a so-called simulcast configuration, i.e. just parallel independent operation of the core coder and the transform coder with the same total bit rate as that of the scalable configuration. First subjective tests using a predecessor of the parametric coder in which the object type for harmonic tones was not yet implemented have shown that this goal can be met [4]. For some input signals which are quite critical for the transform coder, the obtained quality even can exceed the quality obtained by just applying the transform coder at the total bit rate in a non-scalable configuration.

4 Combination with Speech Coding Techniques

Subjective evaluations within the development procedure of the MPEG-4 Audio standard have shown that the coding efficiency of different coding techniques highly depends on the type of coded audio material, if bit rates go down to 16 kbit/s and below. While specialized speech coders show superior performance for speech signals and transform coders have clear advantages for music signals with contents of high complexity, parametric audio coding gives the best quality for music signals with contents of relatively low complexity. For clean speech the parametric speech coding technique Harmonic Vector eXcitation Coding (HVXC) performs at 2 kbit/s clearly better than parametric audio or transform coding at 6 kbit/s. However this highly optimized speech coder is relatively sensitive with respect to background music. In order to gain from the benefits of both parametric approaches it seems useful to investigate in possible combinations with the goal to obtain a system which is capable of encoding speech, music, and mixtures of both.

First experiments have shown that separate encoding of music and speech can improve the performance compared to encoding of the mixture by one single codec. However this procedure is not applicable if speech and music only are available in mixed form. In the following a first concept is presented which allows the automatic assignment of signal components to the speech and audio coder.

Since HVXC is primarily designed for speech signals, it supports coding of a harmonic tone for voiced speech segments or a noise component for unvoiced speech segments. The source model for speech signals is also reflected in the presence of an LPC prefilter which is adapted to the spectral envelope of the input signal. The coding of the parameters is also optimized for coding speech signals, e.g. the amplitudes of the partials in a harmonic tone are vector quantized.

Since there is no possibility to transmit tonal components in addition to the harmonic tone, annoying artifacts can occur if the fundamental frequency estimation (pitch detection) becomes unstable due to the presence of several different harmonic tones or other tonal components added to a clean speech signal.

Therefore the following procedure was developed which ensures that the HVXC receives an input signal only consisting of a single harmonic tone and noise (Figure 5). In the first step the most dominant harmonic tone is extracted by using the tools of the parametric audio encoder. Since HVXC has a more appropriate model for voiced speech segments, the corresponding parameters are not transmitted, but only used for the calculation of a first residual signal. This residual signal is fed to the parametric audio encoder again which now can extract parameters for a second harmonic tone and additional individual sinusoids. The synthesized signal from the first stage is additionally fed to the psychoacoustic model in order to allow proper calculation of masking thresholds for all extracted components. These parameters are encoded and transmitted as with the parametric audio coder in normal operation. The estimation and transmission of parameters for noise components however is disabled. By subtracting the tonal components extracted in this second stage from the original input signal a residual signal is calculated which now contains the first harmonic tone again, but not influenced by modeling errors from the first stage so that this residual signal can be coded by HVXC.

First experiments with the described procedure using mixtures of speech and music showed that the overall stability increases, when compared to the pure speech coder. Even instabilities of the fundamental frequency estimation for the first harmonic tone are not too critical, since the remaining signal is modeled by the harmonic tone and the individual sinusoids of the second stage. Compared to the original parametric audio coder the quality is increased for signals with a dominant speech signal due to the more appropriate model of HVXC. However, if the most dominant signal is a harmonic tone of a musical instrument, the quality can be degraded due to the inappropriate representation by HVXC. Therefore an additional classification seems to be necessary which controls the encoding of the first harmonic tone so that for speech components HVXC can be applied while the technique of the parametric audio coder can be applied to music signals.

5 Conclusions

Two concepts for the combination of parametric audio coding techniques with other coding techniques were presented. The first approach is the combination with transform coding to a scalable system. In comparison to the normal operation of the parametric audio coder, in this configuration additional information needs to be transmitted to enable the calculation of a residual signal in the encoder and the appropriate reconstruction in the decoder. This additional information consists of phase parameters and refinement data for frequency parameters. The resulting scalable coder provides higher efficiency than the two individual coders in simulcast mode and in some cases even performs better than the stand-alone transform coder operating at the same total bit rate.

The second hybrid combination is that of the parametric audio coder and a speech coder. The goal of such a combination is a system which is capable of coding music, speech and mixtures of both. Making use of the object based approach in the parametric audio coder, it is possible to separate the strongest harmonic tone and noise components and to feed them to a speech coder, while all other signal components can be coded using the technique of the parametric audio coder. First results show that improvements can be achieved for speech with background, while further optimization is necessary in order to allow an automatic select of the most suitable coding technique, if the strongest harmonic tone comes from a musical instrument.

References

[1] H. Purnhagen, B. Edler, and C. Ferekidis, "Object-Based Analysis/Synthesis Audio Codec for Very Low Bit Rates," *104th AES Convention*, Preprint 4747, May 1998. `http://www.tnt.uni-hannover.de/project/coding/audio/asac/aes_104.html`

[2] B. Grill et al., "MPEG-4 Audio Committee Draft 14496-3," ISO/IEC JTC1/SC29/WG11 N2203, Mar. 1998.
`http://www.cselt.it/mpeg/standards/mpeg-4.htm`

[3] B. Grill, K. Brandenburg, "MPEG-4 Technical Description - Contribution of Uni Erlangen / FhG-IIS," ISO/IEC JTC1/SC29/WG11 MPEG95/0426, Nov. 1995.

[4] B. Edler, L. Contin, "MPEG-4 Audio Test Results (MOS Test)," ISO/IEC JTC1/SC29/WG11 N1144, Jan. 1996.

[5] M. Nishiguchi et al., "Technical description of Sony IPC's proposals for MPEG-4 audio and speech coding (revised)," ISO/IEC JTC1/SC29/WG11 MPEG96/0731, Mar. 1996.

[6] H. Purnhagen et al., "Proposal for the Integration of Parametric Speech and Audio Coding Tools based on an Automatic Speech/Music Classification Tool," ISO/IEC JTC1/SC29/WG11 MPEG97/2481, Jul. 1997.

[7] C. Ferekidis, *Kriterien für die Auswahl von Spektralkomponenten in einem Analyse/Synthese-Audiocoder*. Diplomarbeit (diploma thesis), Universität Hannover, Institut für Theoretische Nachrichtentechnik und Informationsverarbeitung, Mar. 1997.

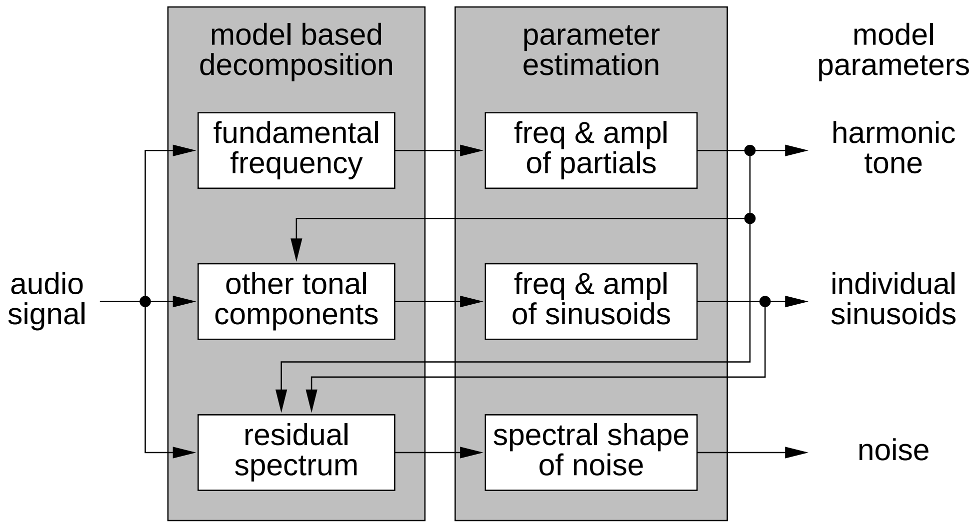

Figure 1: Signal decomposition and parameter estimation for the extended source model utilising the object models *harmonic tone*, *individual sinusoid*, and *noise*.

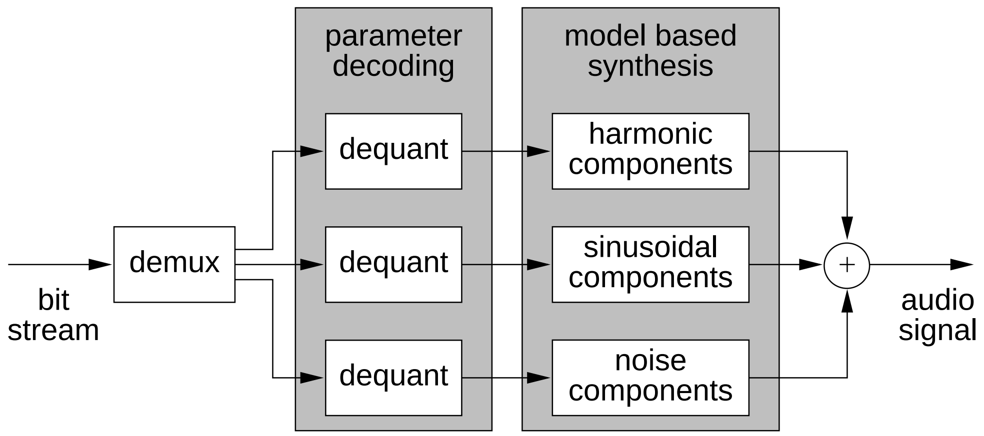

Figure 2: Block diagram of an object-based audio decoder.

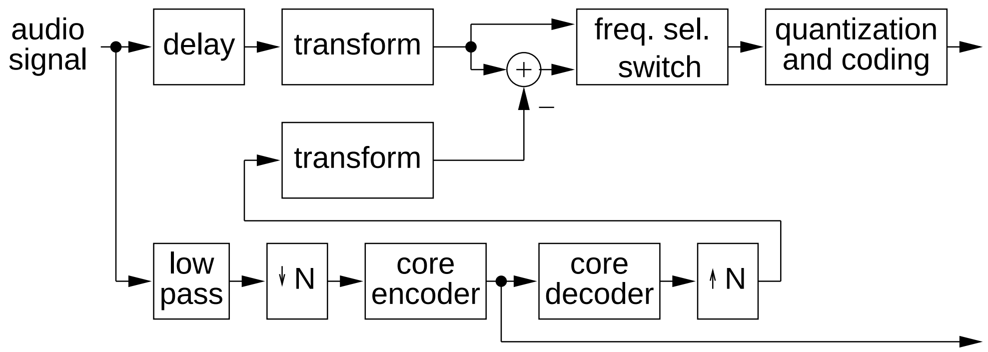

Figure 3: Scalability tools for combining a transform coder with a low rate core coder.

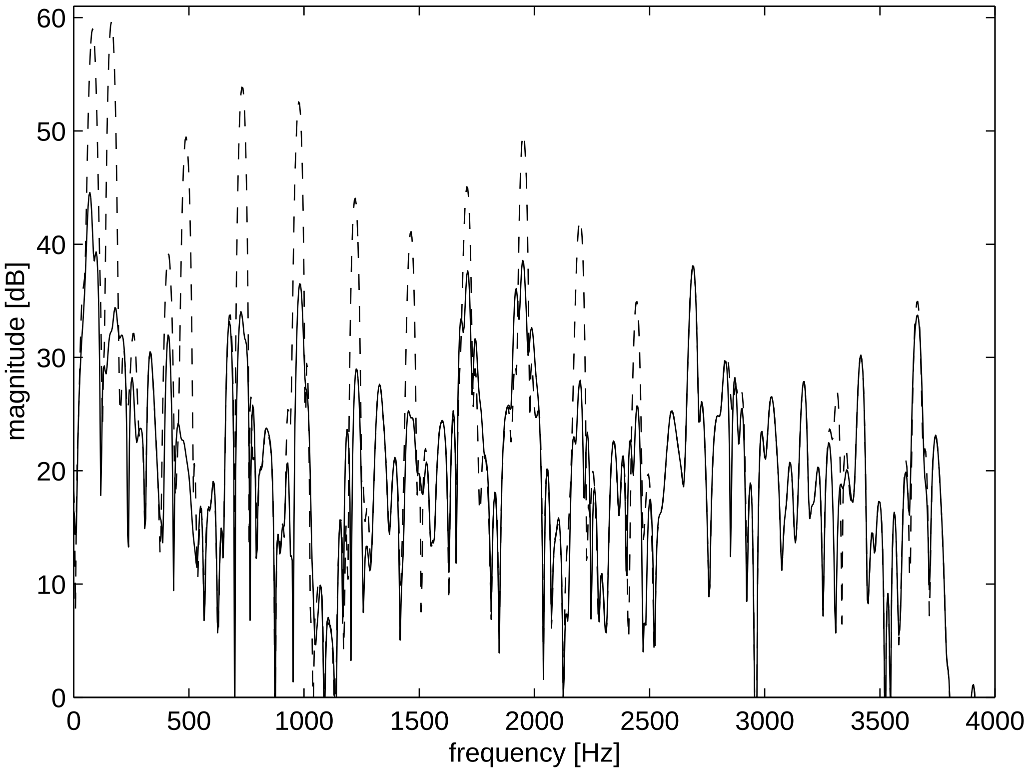

Figure 4: Short time spectra of original signal (- -) and residual (—) for an excerpt of test signal "eddie rabbit".

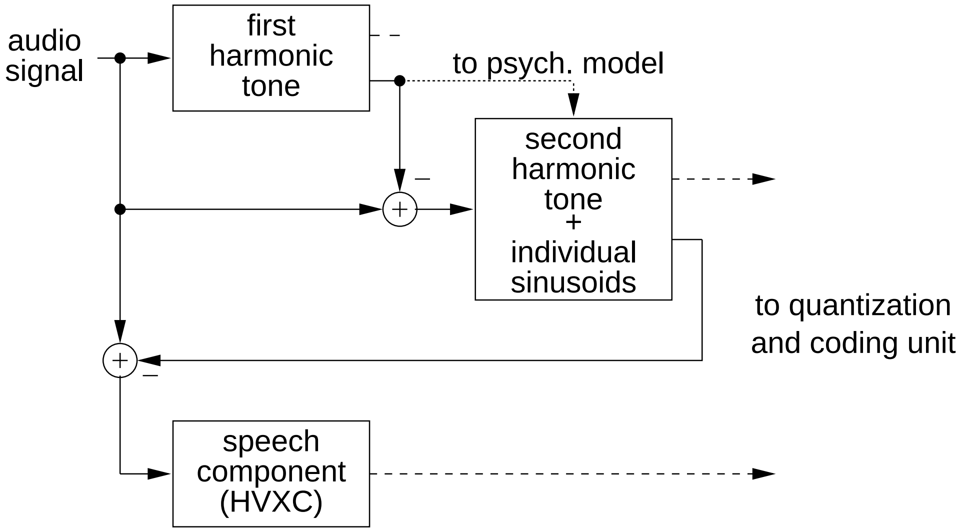

Figure 5: Extended signal decomposition for speech and music (signal samples: —, parameters: - -).